A practical approach to working with the BQC19 multi-*omics datasets

- Approaching a new type of data -

Antoine Soulé,
Université McGill, Montréal, QC, Canada

MUHC

29 Novembre 2022

antoine.soule@mcgill.ca

SomaScan

A new tool

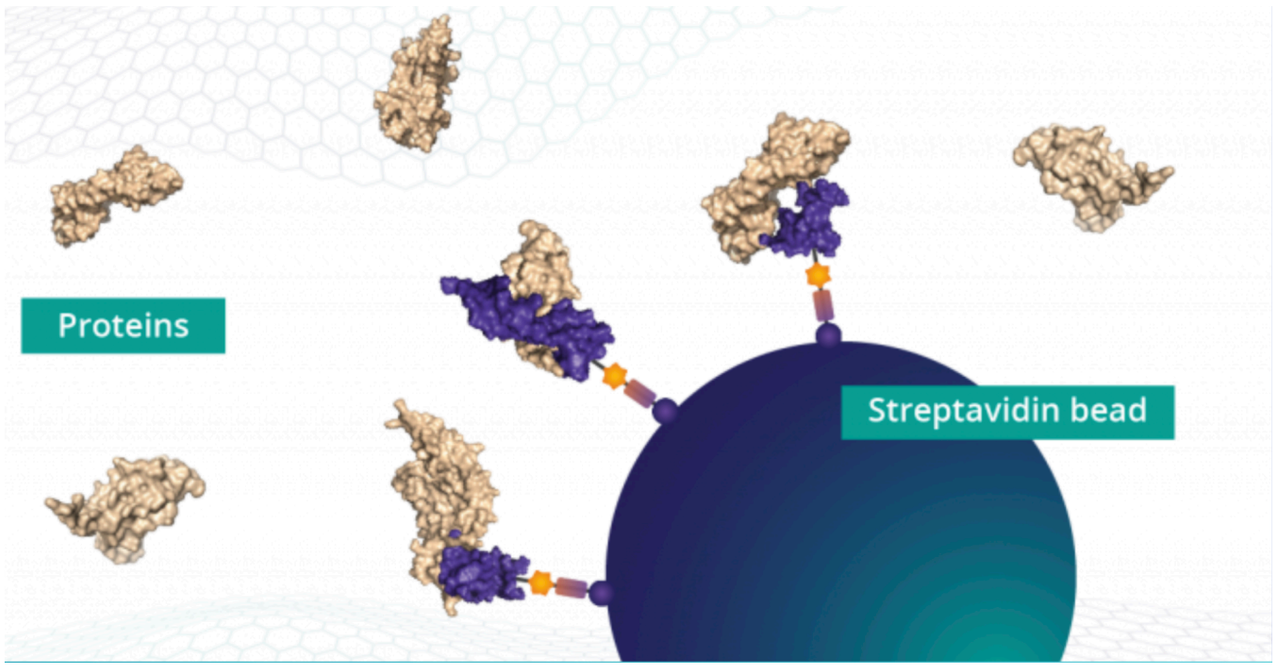
Rely on **aptamers**

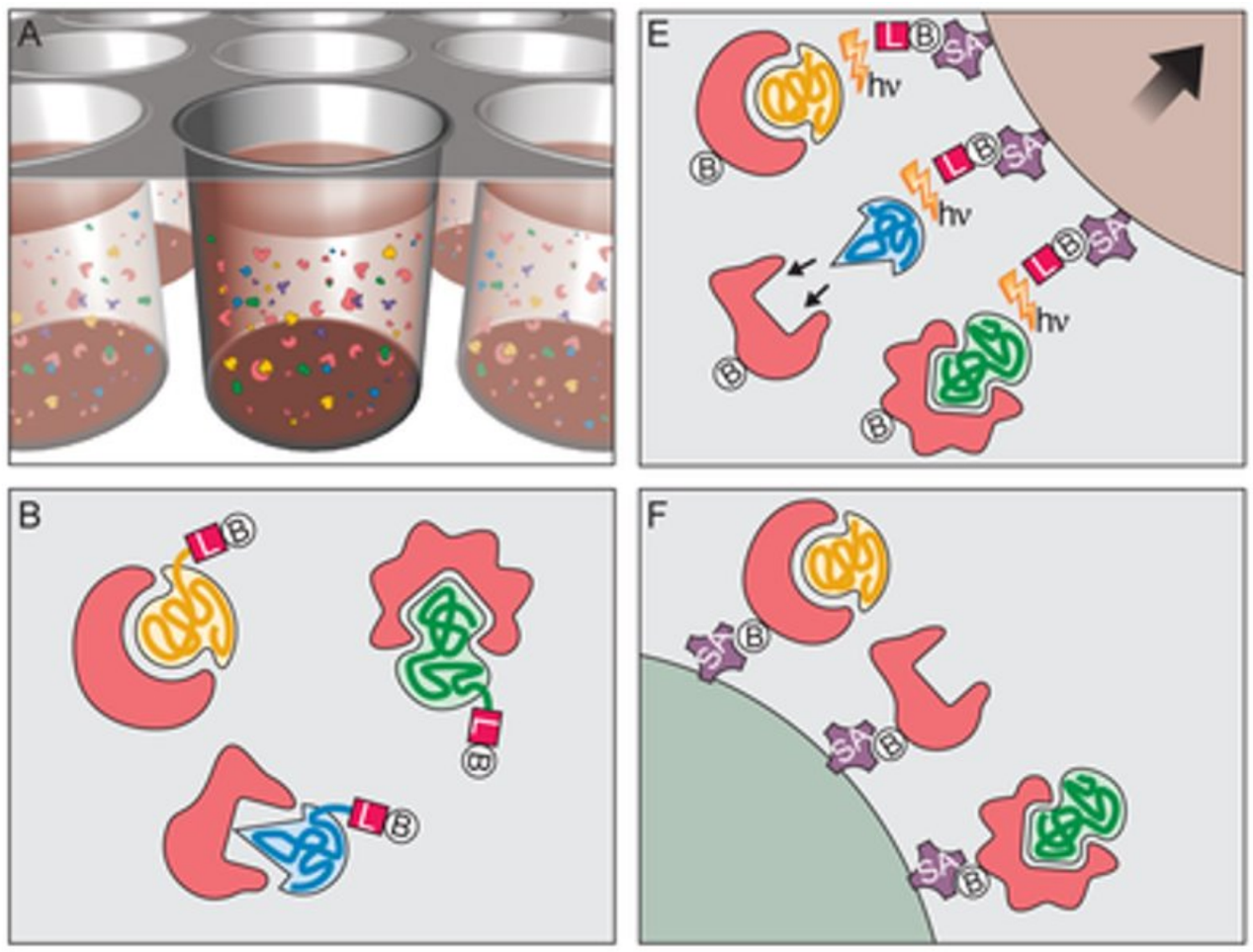
Linked to **flurophores**

To *measure* the expression of ~5K proteins

relatives measurements ⇒ normalisation & co

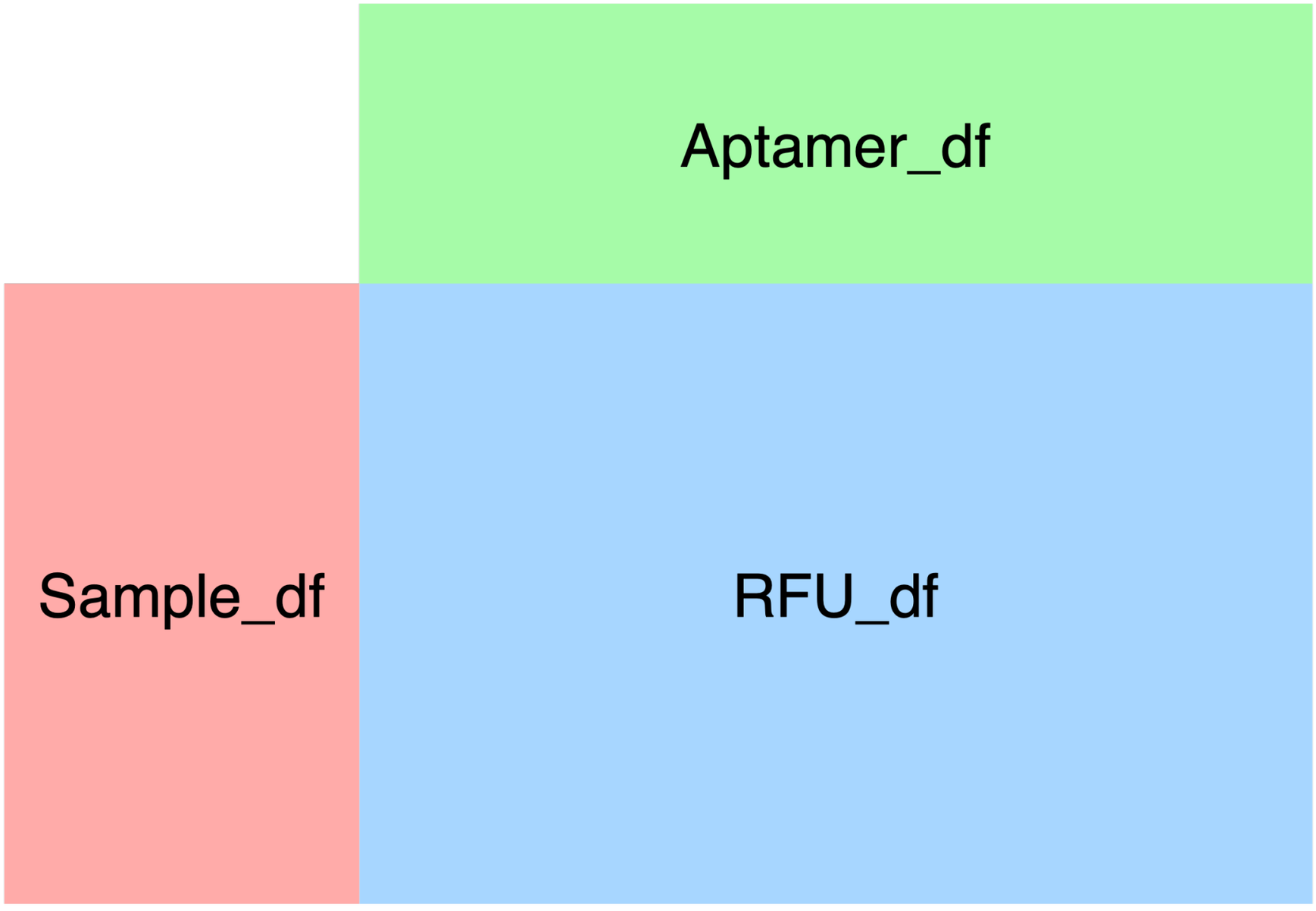
“What I wish I knew when I started using those data”





SomaScan - The Adat format

3 DataFrames in 1

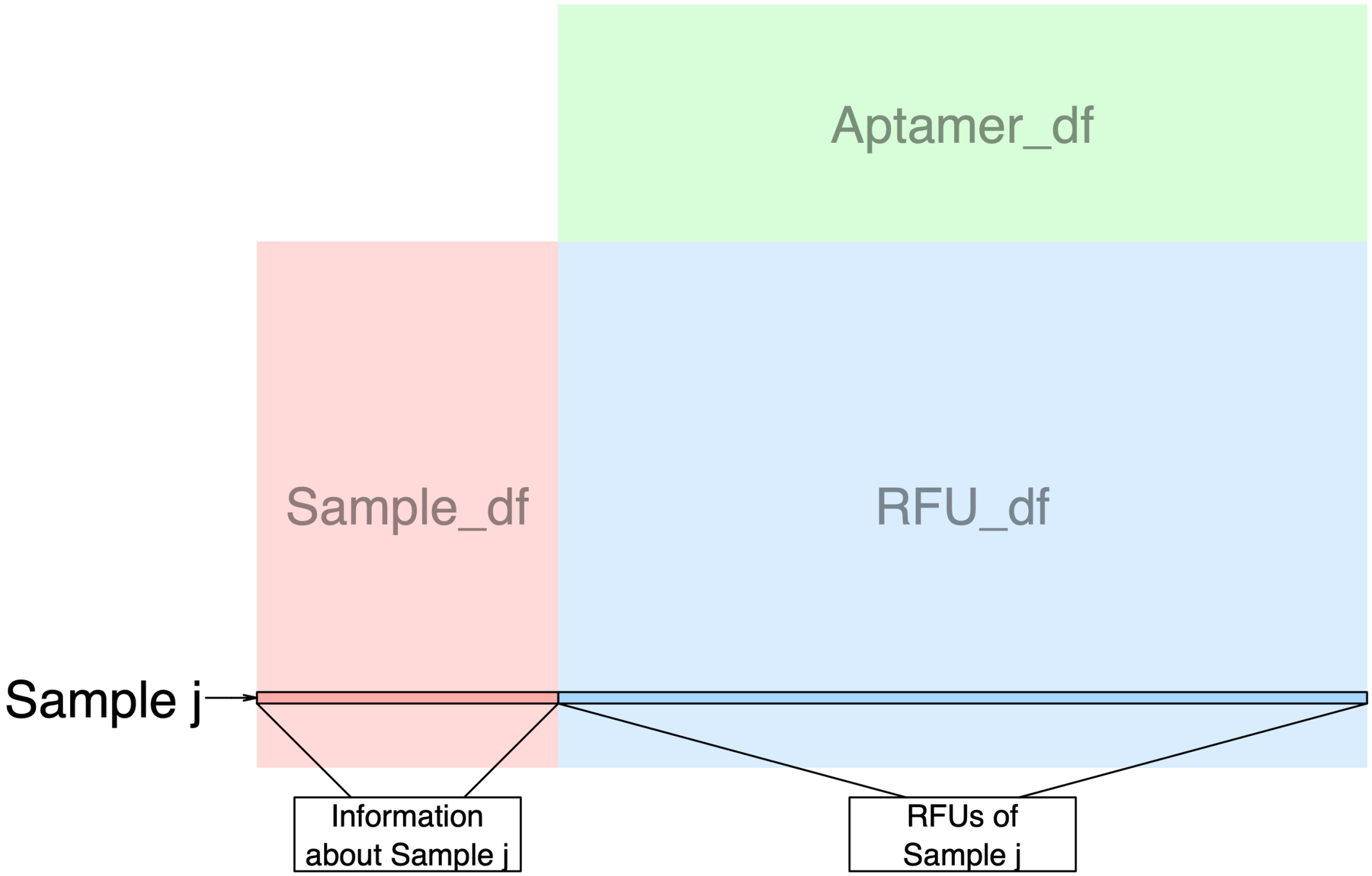



```
import canopy

adat = canopy.read_adat(file_path)
```

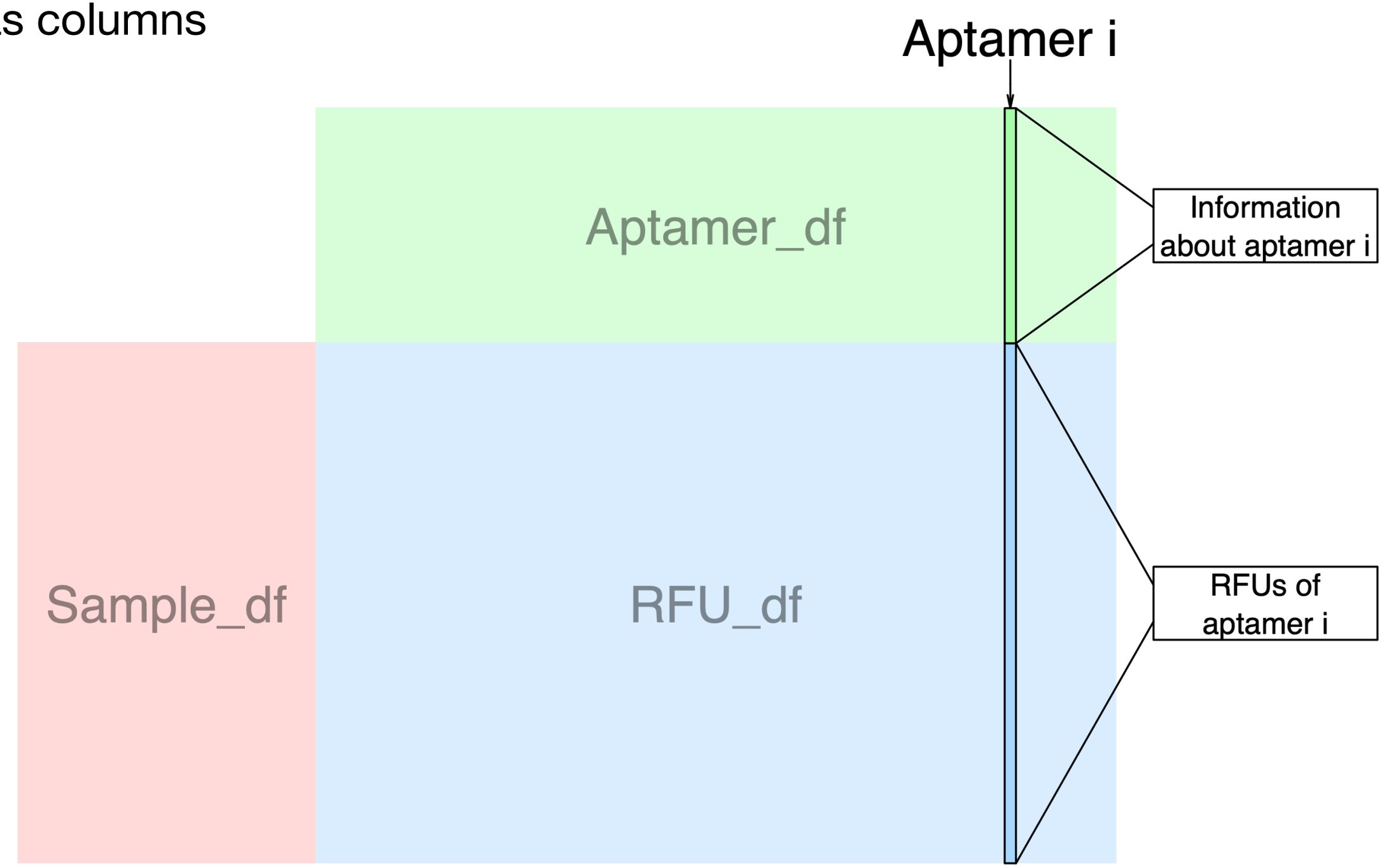
SomaScan - The Adat format

Samples as Rows



SomaScan - The Adat format

Aptamers as columns

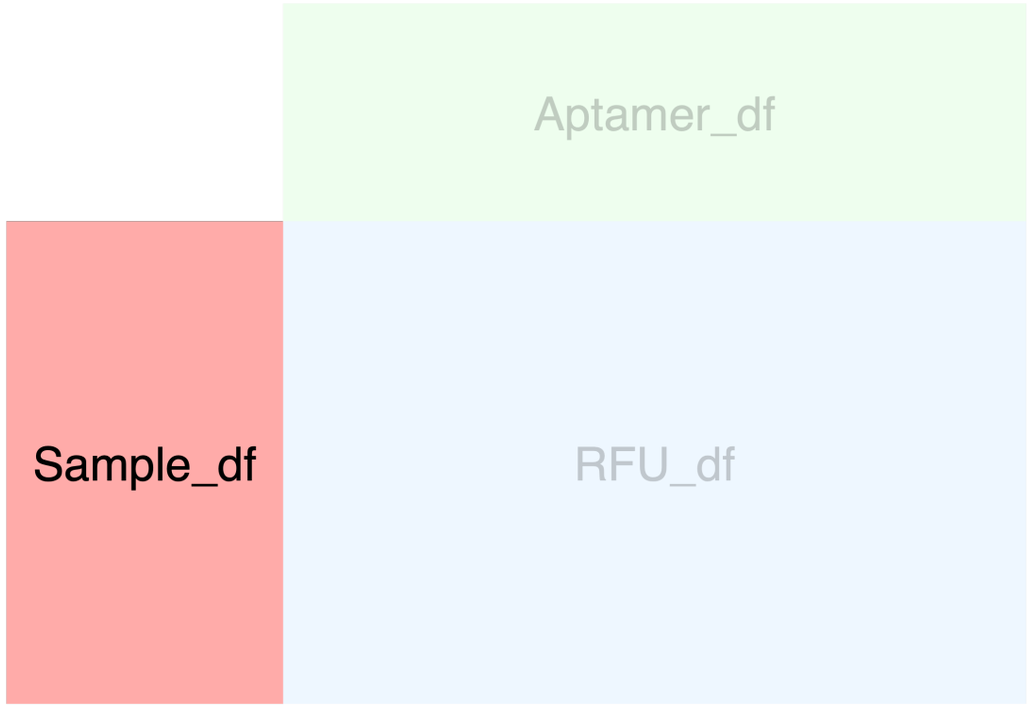


SomaScan - The Adat format

Information on samples unrelated to aptamers

Sample_df

- **SubjectID** (VAP12345-78)
- PlateId, PlateRunDate, etc,…
- RowCheck (PASS/FLAG ?)
- +~30 others…



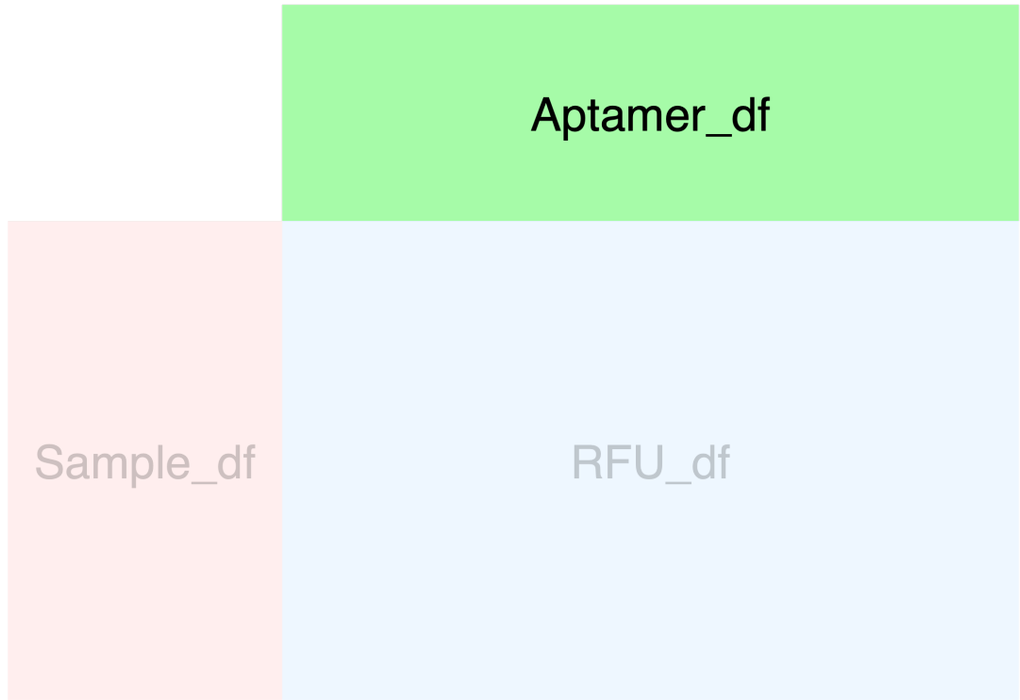
sample_df = adat.index.to_frame(index=False)

SomaScan - The Adat format

Information on aptamers unrelated to samples

Aptamer_df `< MultiIndex >`

- Aptamer ID (`SomaId`, unique)
- Protein IDs
  - Uniprot
  - EntrezGeneID
- Calibration values
- Dilution ratios
- Etc,..



WARNING#1:
1 aptamer → 1 protein
1 protein → X aptamer(s)

Different affinities to different protein conformations / states

aptamer_df = adat.columns.to_frame(index=False)

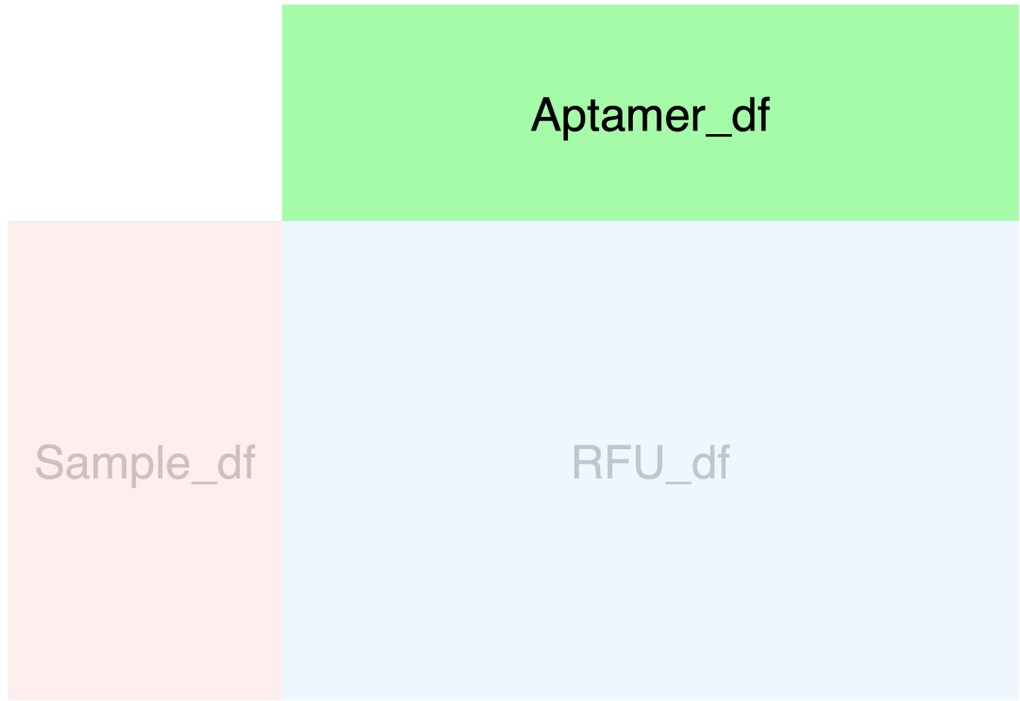
SomaScan - The Adat format

Information on aptamers unrelated to samples

Aptamer_df `< MultiIndex >`

- Aptamer ID (`SomaId`, unique)
- Protein IDs
  - Uniprot
  - EntrezGeneID

- Calibration values
- Dilution ratios
- Etc,..



WARNING #2:
We **shouldn't** compare RFUs of **aptamer A and B** in patient X
We **can** compare RFUs of aptamer A between **patient X and Y**

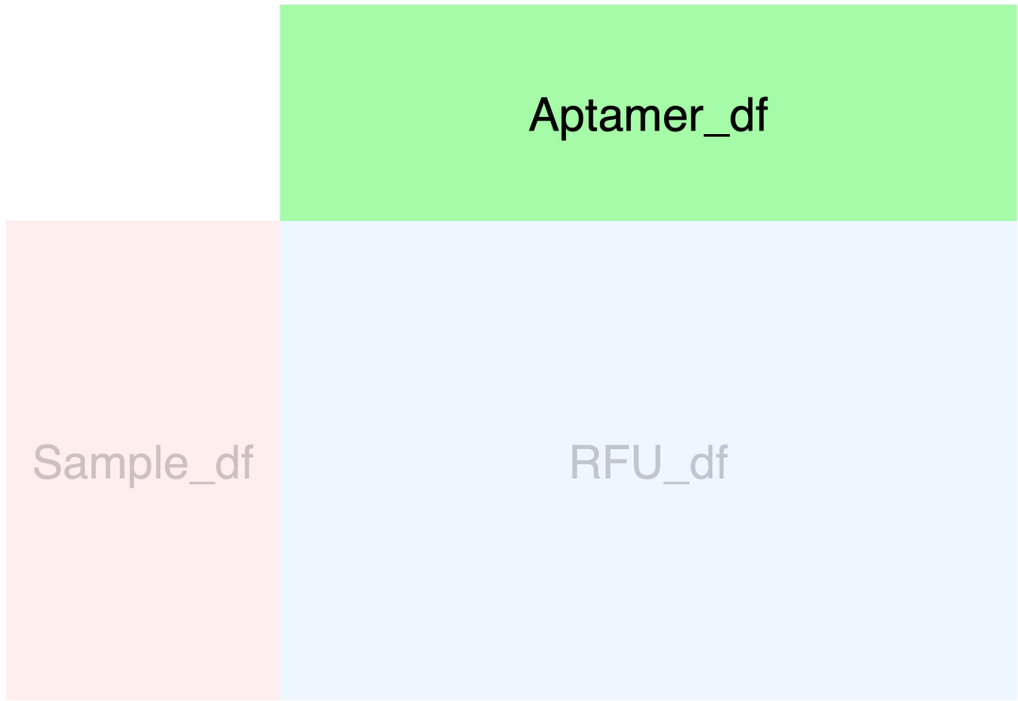
aptamer_df = adat.columns.to_frame(index=False)

SomaScan - The Adat format

Information on aptamers unrelated to samples

Aptamer_df `< MultiIndex >`

- Aptamer ID (`SomaId`, unique)
- Protein IDs
  - Uniprot
  - EntrezGeneID
- Calibration values
- Dilution ratios
- Etc,..



WARNING #3: Includes control aptamers !
You might want to filter:
- ‘Organism’ == ‘Human’
- ‘Type’ == ‘Protein’

5284 → 4979 aptamers

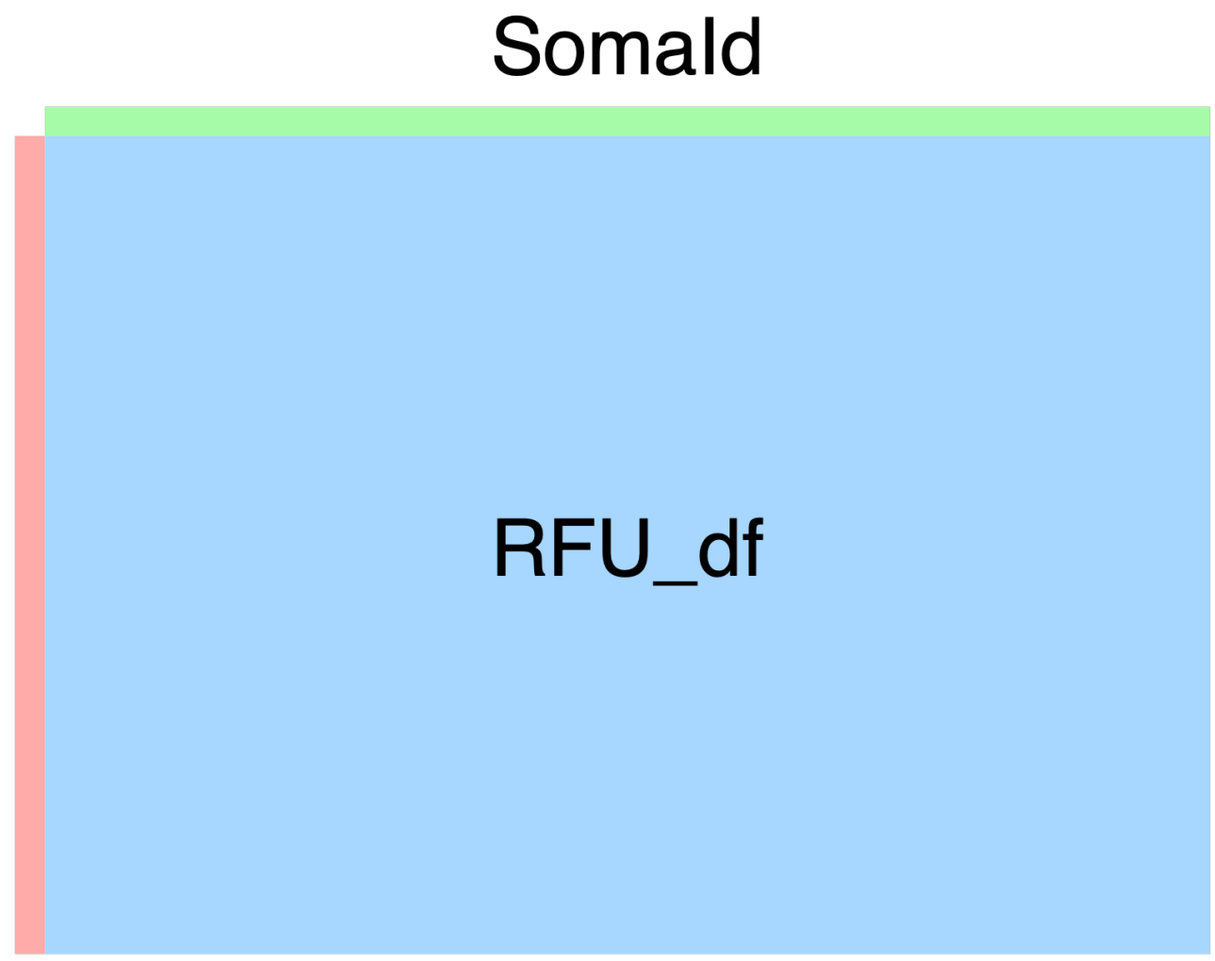
SomaScan - The Adat format

Down to the essential…

Simplest solution IMO:
cut it back to a standard Pandas.DataFrame

RFU_df

- Sample_df → SubjectID *index*
- Aptamer_df → SomaId *columns*



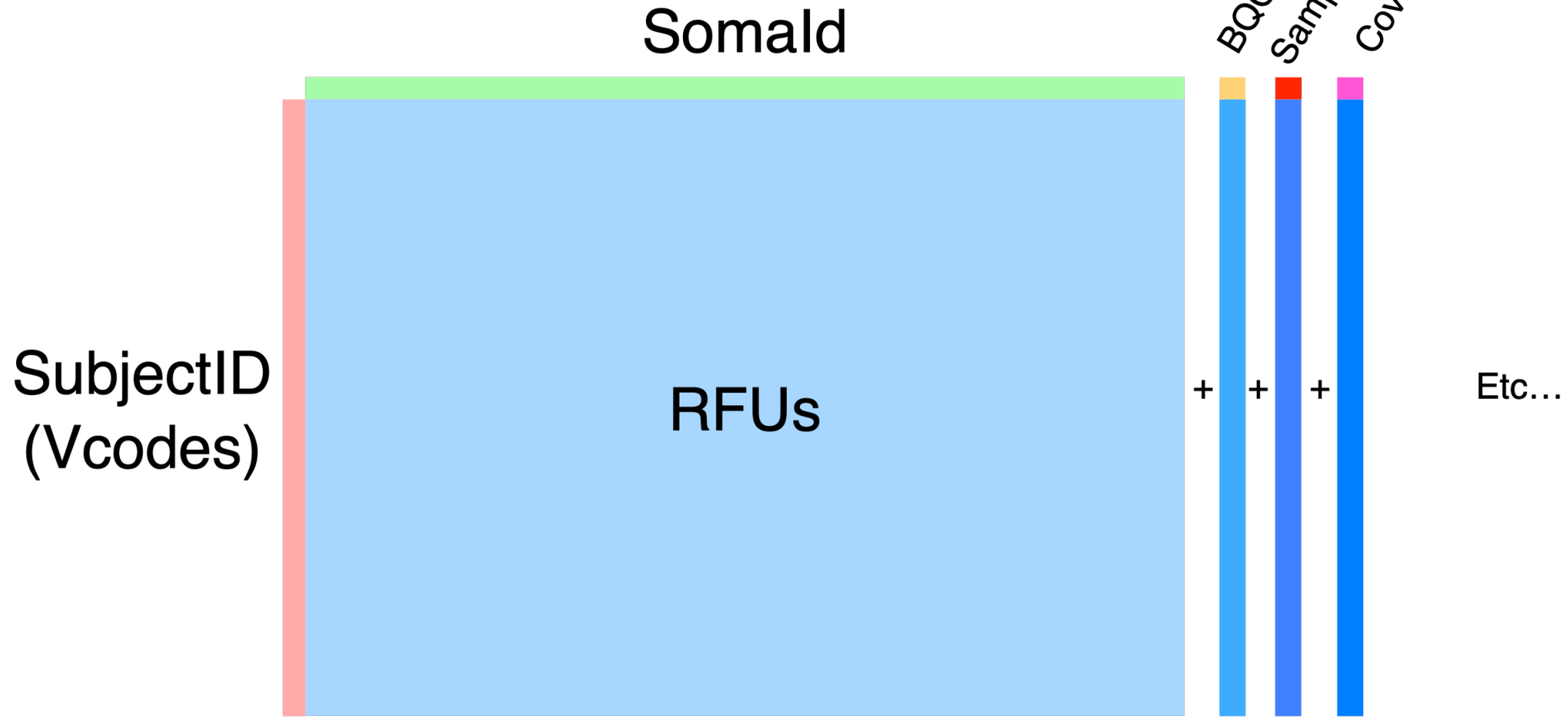

```
df_RFU=adat.pick_on_meta(axis=0, name='SampleType', values=['Sample']).pick_meta(axis=0, names=['SubjectID'])
df_RFU.columns=df_RFU.columns.get_level_values('SomaId')
```

SomaScan - The Adat format

…and add whatever you need.

Then you can:

- Use aptamer_df to determine which SomaIds are interesting
- Use the Vcode mapping file to map Vcodes to BQC_IDs, dates, etc,…
  - with BQC_IDs + dates you can map even more info
  - → New columns → filters → drop new columns



SomaScan - 1 dataset but 4 files

Couldn’t be that simple, right ?

There is actually 4 adat files available in BQC19 because:

1. Samples were processed in two separate batches
2. Each batch is available raw or already normalised by Somalogic

So 2 batches x 2 versions = 4 files

Our two cents:

Somalogic’s normalisation method seemed sound so
we used for the normalised versions

We log2 and z-score normalised (on columns) each batch separately
before concatenating them to reduce batch effect.

Unsupervised Clustering

One way to use those data

Idea: comparing “mild” cases vs.“severe”

mixes together des heterogeneous profiles

which blurs any signal

Ideally, we would like to separate those profiles

and study their differences

↓

Unsupervised clustering of patients

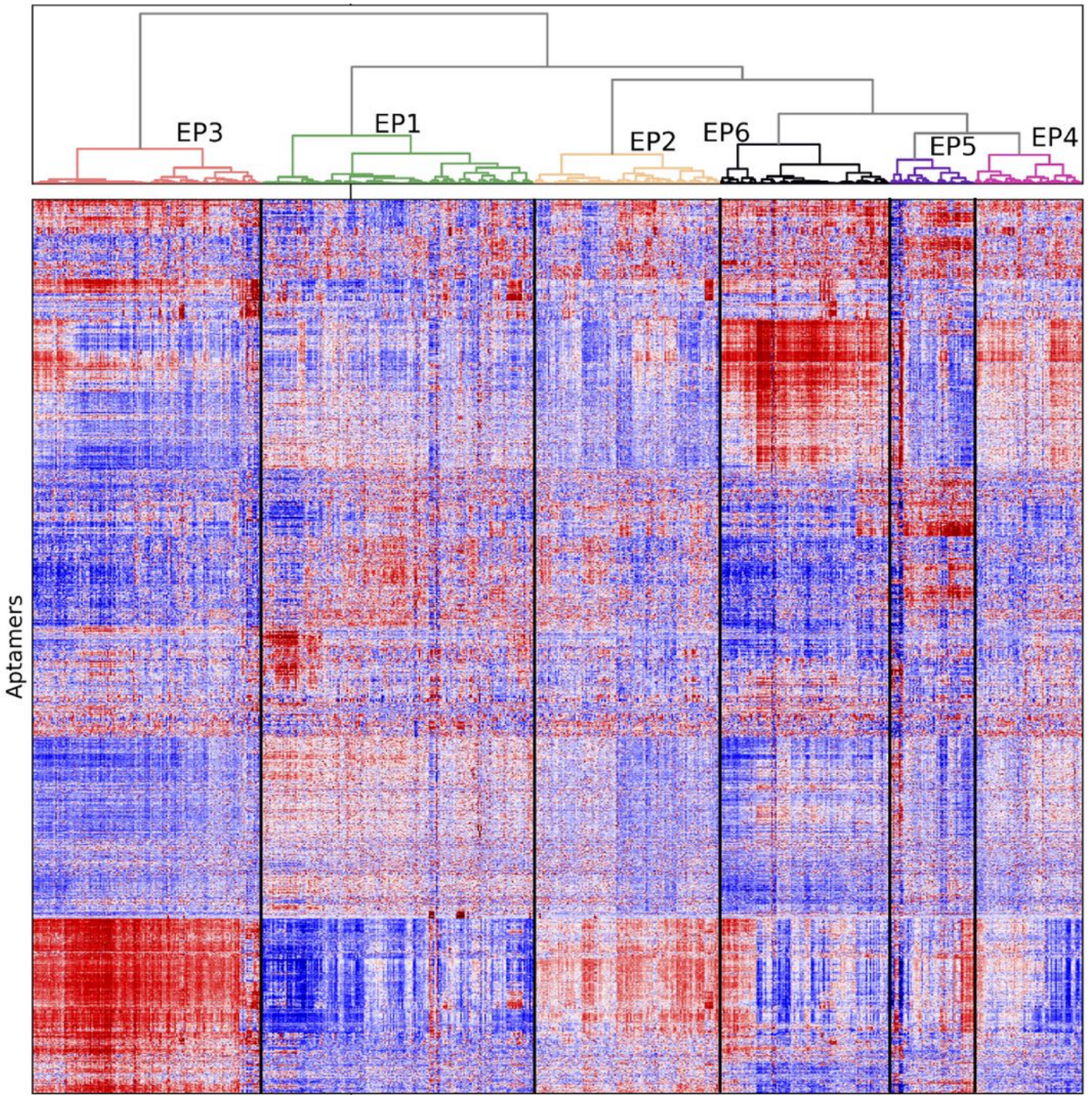
based on their protein expression

only

(First blood draw)

731 patients

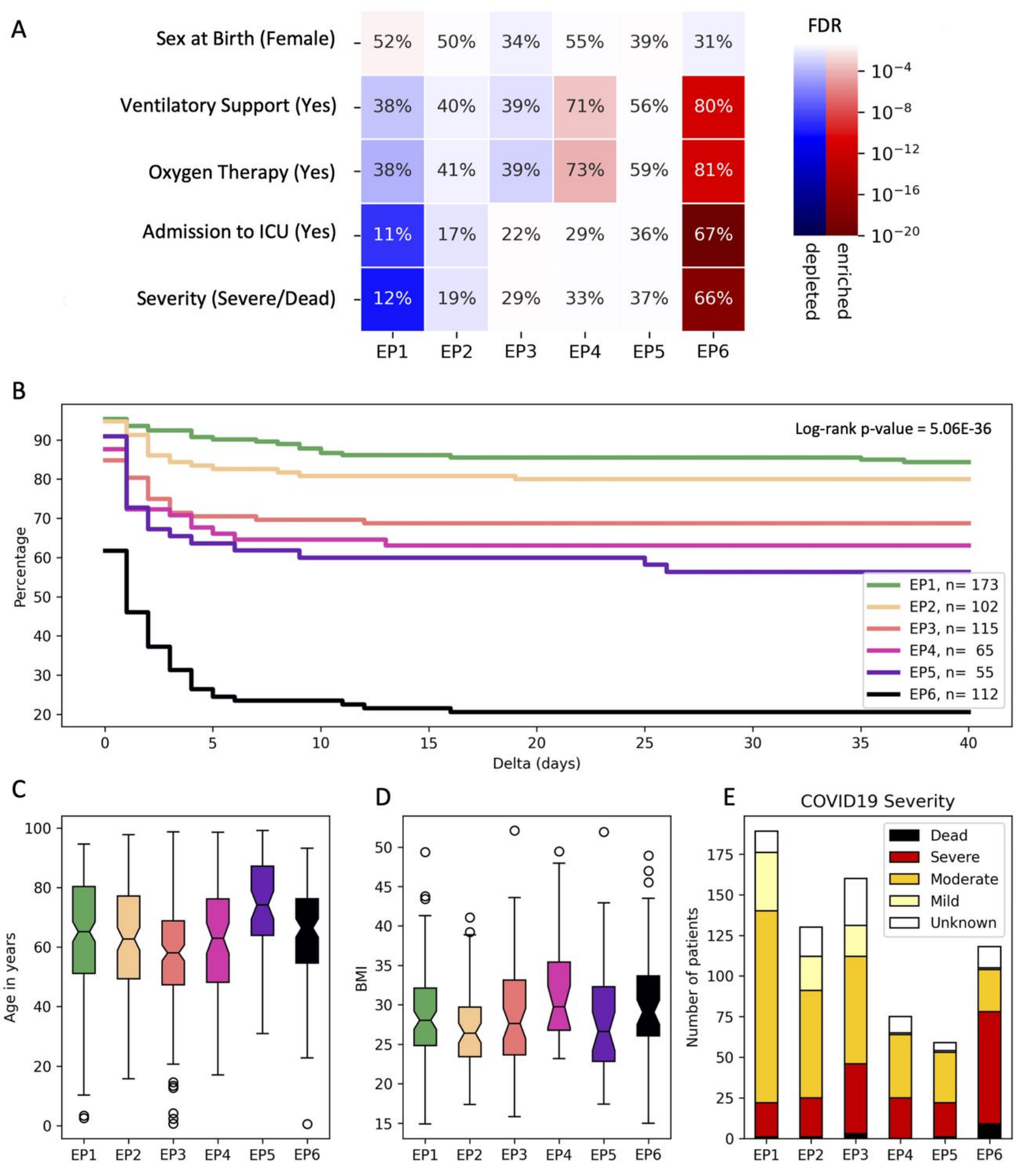
Hospitalised **and** COVID positif



Unsupervised clustering of SARS-CoV-2 hospitalized patients identifies FGFR-signaling in severe COVID-19 acute respiratory distress syndrome

EP

Characterisation



EP

Characterisation

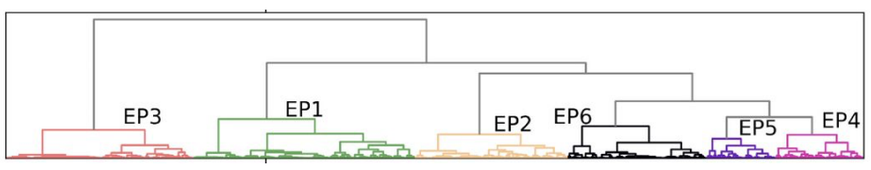
Kaplan-Meier

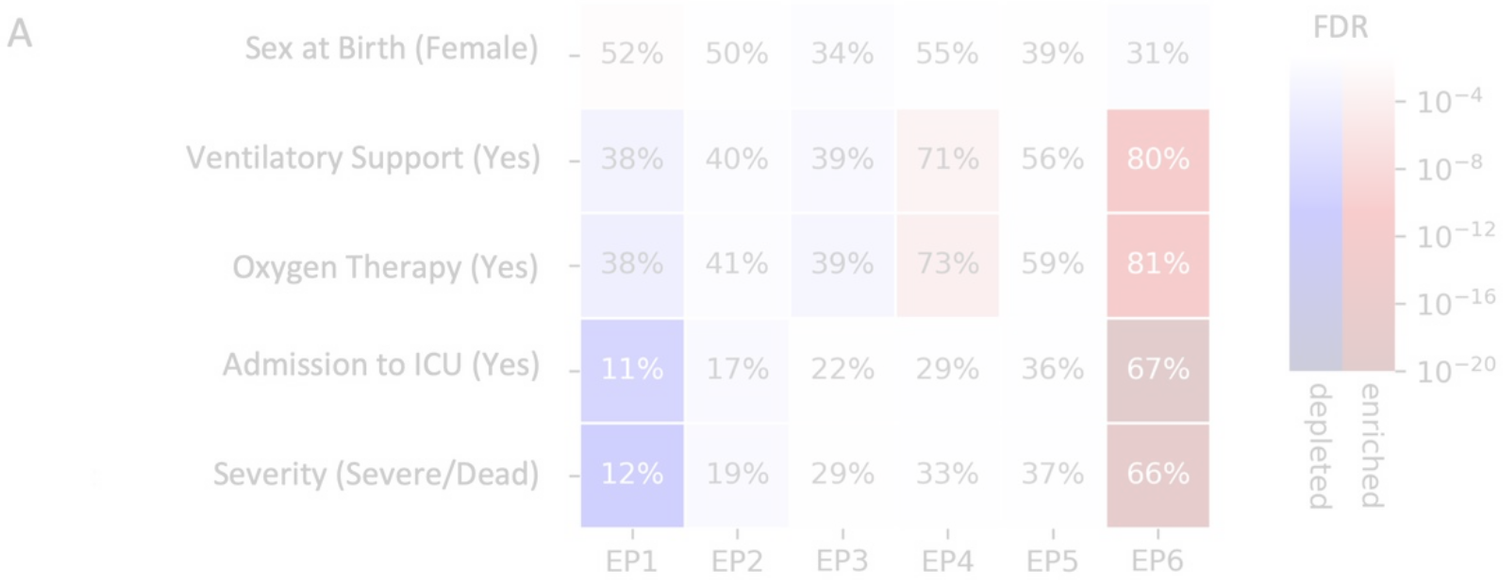
D0= admission to hospital

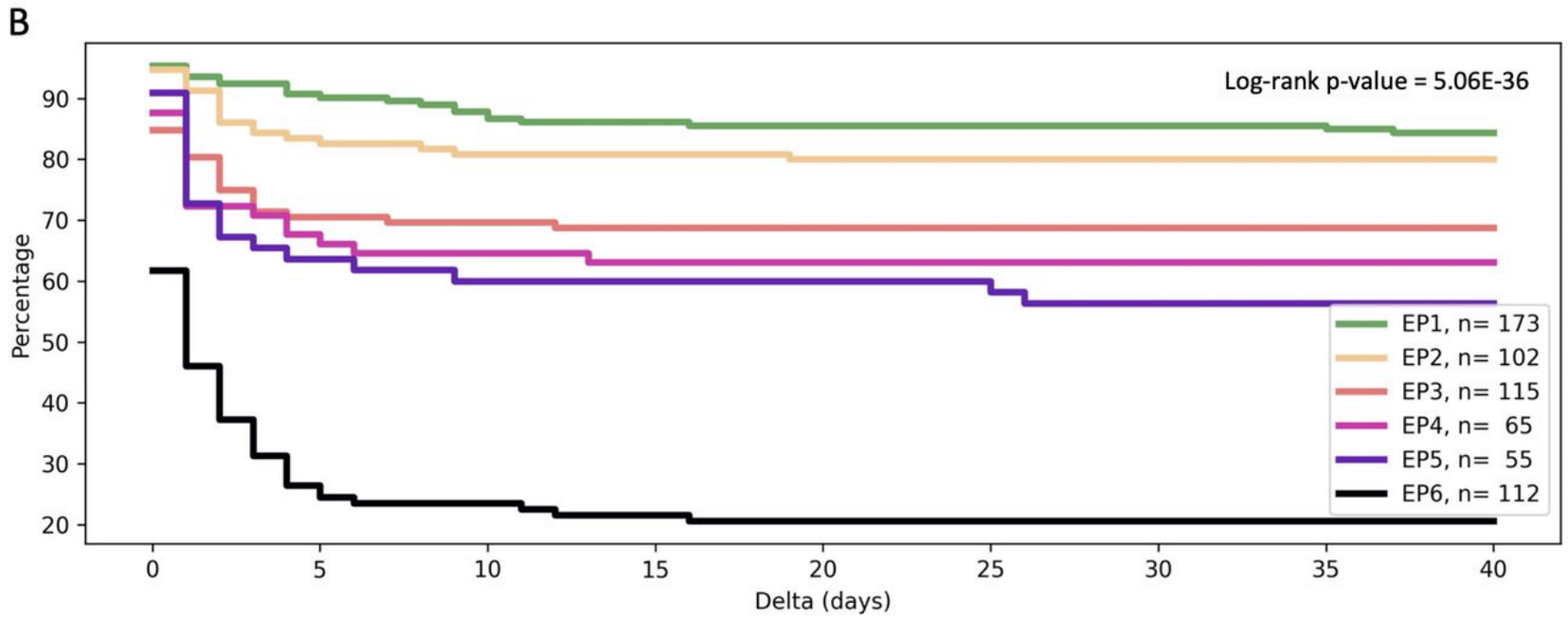
E= admission to ICU (or death)

Clear deterioration of prognostic

⇒ EPs' numbering







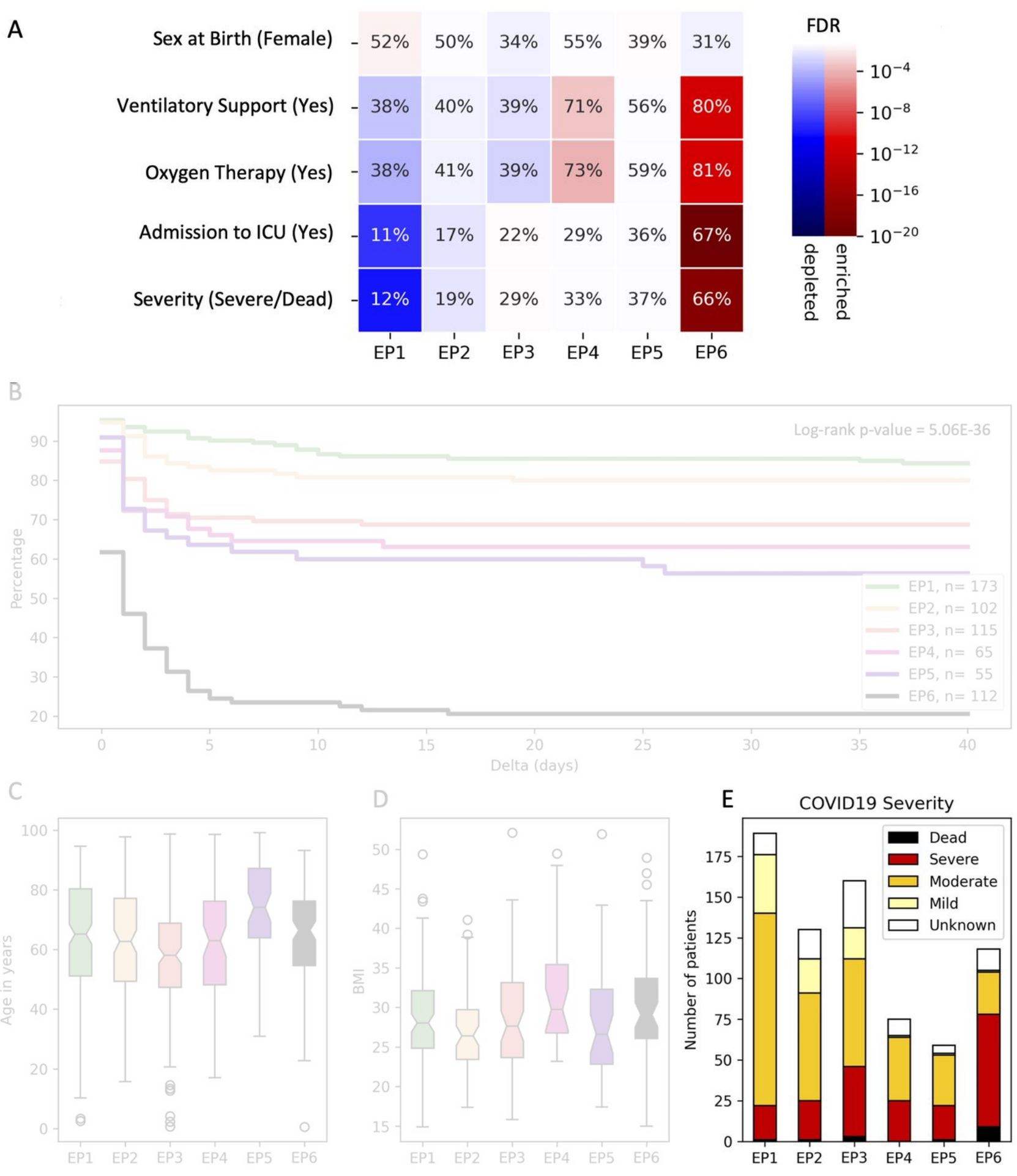
EP

Characterisation

Clear deterioration of prognostic

EP1 and **EP6** as the

two opposite extremes



EP

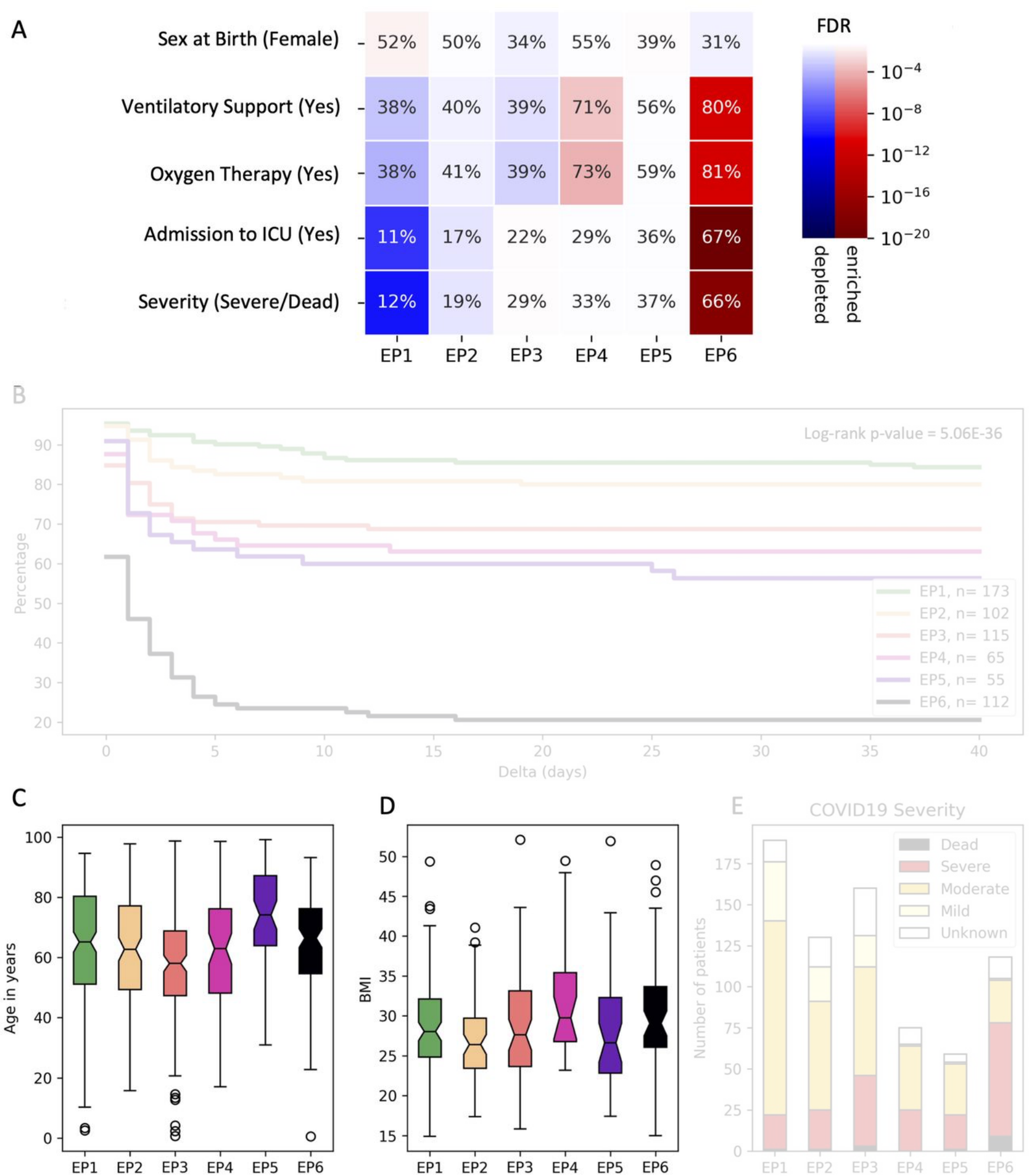
Characterisation

Weak connection with known suspects:

Sex: weak correlation (cf. EP3)

Age: also weak (cf. EP5)

BMI: unclear (missing data)



Thanks for your attention,

For the invitation,

And for the data !

Questions are obviously welcome !

antoine.soule@mcgill.ca